CS 412 Assignment 3 Report

Guangyu Zhou

Language chosen: C++

Step 1:

***Question to ponder A****: How do you choose min\_sup for this task? Explain how you choose the min\_sup in your report. Any reasonable choice will be fine.*

***Answer: I choose min\_sup = 100. The reason is that for this min\_sup, we have about 100 frequent patterns, which is about 1% of the original dataset. From these patterns, we can find enough key frequent word and phrases related to the topic.***

Step 2:

***Question to ponder B:*** *Can you figure out which topic corresponds to which domain based on patterns you mine? Write your observations in the report.*

***Answer: Here is my observation: Basically I look for the most frequent top 10 patterns for each topic and matching them with the relevant name in the field.***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Topic 0* | *Topic 1* | *Topic 2* | *Topic 3* | *Topic 4* |
| *Data mining* | *Machine learning* | *Information Retrieval* | *Theory* | *Database* |

***Question to ponder C:*** *Compare the result of frequent patterns, maximal patterns and closed patterns, is the result satisfying? Write down your analysis.*

***Answer: There are less patterns of maximal patterns than the frequent patterns because of the deletion of those unsatisfied superset patterns. But the closed patterns are basically the same as the frequent patterns. So it may be the chosen of minimal support is too high that there are few patterns with more than one elements.***

Step 3:

***Question to ponder D****: What are the quality of the phrases that satisfies both min\_sup and min\_conf? Please compare it to the results of Step1 and put down your observations.*

***Answer: The key difference between the result generated by Weka and my FP growth tree in step 1 is that this part involves min confidence. As a result, to satisfy both the min support and min confidence, some patterns with high support only, like “data”, are not included in the Weka result because they have low confidence.***

|  |  |
| --- | --- |
| Topic 0  Min support = 50  Min confi = 0.5 | === Run information ===  Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 50.0  Relation: topic-0.txt  Instances: 10047  Attributes: 134  [list of attributes omitted]  === Associator model (full training set) ===  FPGrowth found 13 rules (displaying top 10)  1. [series=1]: 209 ==> [time=1]: 194 <conf:(0.93)> lift:(16.65) lev:(0.02) conv:(12.33)  2. [lower=1]: 63 ==> [bound=1]: 57 <conf:(0.9)> lift:(49.95) lev:(0.01) conv:(8.84)  3. [mining=1, rule=1]: 159 ==> [association=1]: 123 <conf:(0.77)> lift:(23.13) lev:(0.01) conv:(4.15)  4. [mining=1, association=1]: 159 ==> [rule=1]: 123 <conf:(0.77)> lift:(18.68) lev:(0.01) conv:(4.12)  5. [dimensionality=1]: 77 ==> [reduction=1]: 56 <conf:(0.73)> lift:(50.39) lev:(0.01) conv:(3.45)  6. [association=1]: 336 ==> [rule=1]: 233 <conf:(0.69)> lift:(16.75) lev:(0.02) conv:(3.1)  7. [stream=1]: 211 ==> [data=1]: 141 <conf:(0.67)> lift:(5.21) lev:(0.01) conv:(2.59)  8. [decision=1]: 113 ==> [tree=1]: 74 <conf:(0.65)> lift:(12.93) lev:(0.01) conv:(2.68)  9. [rule=1]: 416 ==> [association=1]: 233 <conf:(0.56)> lift:(16.75) lev:(0.02) conv:(2.19)  10. [frequent=1]: 227 ==> [mining=1]: 127 <conf:(0.56)> lift:(4.83) lev:(0.01) conv:(1.99) |
| Topic 1  Min support = 50  Min confi = 0.5 | === Run information ===  Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 50.0  Relation: topic-1.txt  Instances: 9674  Attributes: 126  [list of attributes omitted]  === Associator model (full training set) ===  FPGrowth found 19 rules (displaying top 10)  1. [machine=1, support=1]: 117 ==> [vector=1]: 115 <conf:(0.98)> lift:(42.26) lev:(0.01) conv:(38.09)  2. [machine=1, vector=1]: 123 ==> [support=1]: 115 <conf:(0.93)> lift:(41.68) lev:(0.01) conv:(13.36)  3. [learning=1, semi=1]: 55 ==> [supervised=1]: 51 <conf:(0.93)> lift:(50.97) lev:(0.01) conv:(10.8)  4. [neighbor=1]: 137 ==> [nearest=1]: 121 <conf:(0.88)> lift:(60.6) lev:(0.01) conv:(7.94)  5. [nearest=1]: 141 ==> [neighbor=1]: 121 <conf:(0.86)> lift:(60.6) lev:(0.01) conv:(6.62)  6. [neural=1]: 128 ==> [network=1]: 101 <conf:(0.79)> lift:(16.49) lev:(0.01) conv:(4.35)  7. [vector=1, support=1]: 146 ==> [machine=1]: 115 <conf:(0.79)> lift:(24.66) lev:(0.01) conv:(4.42)  8. [support=1]: 217 ==> [vector=1]: 146 <conf:(0.67)> lift:(28.93) lev:(0.01) conv:(2.94)  9. [vector=1]: 225 ==> [support=1]: 146 <conf:(0.65)> lift:(28.93) lev:(0.01) conv:(2.75)  10. [learning=1, supervised=1]: 80 ==> [semi=1]: 51 <conf:(0.64)> lift:(37.38) lev:(0.01) conv:(2.62) |
| Topic 2  Min support = 50  Min confi = 0.5 | === Run information ===  Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 50.0  Relation: topic-2.txt  Instances: 9959  Attributes: 141  [list of attributes omitted]  === Associator model (full training set) ===  FPGrowth found 10 rules (displaying top 10)  1. [answering=1]: 88 ==> [question=1]: 77 <conf:(0.88)> lift:(72.02) lev:(0.01) conv:(7.24)  2. [page=1]: 131 ==> [web=1]: 107 <conf:(0.82)> lift:(6.63) lev:(0.01) conv:(4.59)  3. [natural=1]: 228 ==> [language=1]: 170 <conf:(0.75)> lift:(15.15) lev:(0.02) conv:(3.67)  4. [information=1, language=1]: 75 ==> [retrieval=1]: 55 <conf:(0.73)> lift:(6.56) lev:(0) conv:(3.17)  5. [retrieval=1, language=1]: 77 ==> [information=1]: 55 <conf:(0.71)> lift:(5.87) lev:(0) conv:(2.94)  6. [site=1]: 94 ==> [web=1]: 65 <conf:(0.69)> lift:(5.62) lev:(0.01) conv:(2.75)  7. [engine=1]: 181 ==> [search=1]: 122 <conf:(0.67)> lift:(9.49) lev:(0.01) conv:(2.8)  8. [question=1]: 121 ==> [answering=1]: 77 <conf:(0.64)> lift:(72.02) lev:(0.01) conv:(2.67)  9. [result=1]: 120 ==> [search=1]: 63 <conf:(0.53)> lift:(7.4) lev:(0.01) conv:(1.92)  10. [retrieval=1, system=1]: 147 ==> [information=1]: 76 <conf:(0.52)> lift:(4.25) lev:(0.01) conv:(1.79) |
| Topic 3  Min support = 50  Min confi = 0.5 | === Run information ===  Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 50.0  Relation: topic-3.txt  Instances: 10161  Attributes: 145  [list of attributes omitted]  === Associator model (full training set) ===  FPGrowth found 11 rules (displaying top 10)  1. [database=1, oriented=1]: 75 ==> [object=1]: 63 <conf:(0.84)> lift:(38.27) lev:(0.01) conv:(5.64)  2. [satisfaction=1]: 96 ==> [constraint=1]: 80 <conf:(0.83)> lift:(19.88) lev:(0.01) conv:(5.41)  3. [artificial=1]: 79 ==> [intelligence=1]: 65 <conf:(0.82)> lift:(73.34) lev:(0.01) conv:(5.21)  4. [database=1, object=1]: 84 ==> [oriented=1]: 63 <conf:(0.75)> lift:(46.75) lev:(0.01) conv:(3.76)  5. [expert=1]: 125 ==> [system=1]: 82 <conf:(0.66)> lift:(7.18) lev:(0.01) conv:(2.58)  6. [oriented=1]: 163 ==> [object=1]: 102 <conf:(0.63)> lift:(28.51) lev:(0.01) conv:(2.57)  7. [deductive=1]: 85 ==> [database=1]: 53 <conf:(0.62)> lift:(5.9) lev:(0) conv:(2.3)  8. [object=1, oriented=1]: 102 ==> [database=1]: 63 <conf:(0.62)> lift:(5.84) lev:(0.01) conv:(2.28)  9. [intelligence=1]: 114 ==> [artificial=1]: 65 <conf:(0.57)> lift:(73.34) lev:(0.01) conv:(2.26)  10. [reinforcement=1]: 224 ==> [learning=1]: 117 <conf:(0.52)> lift:(9.51) lev:(0.01) conv:(1.96) |
| Topic 4  Min support = 50  Min confi = 0.5 | === Run information ===  Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 50.0  Relation: topic-4.txt  Instances: 9845  Attributes: 132  [list of attributes omitted]  === Associator model (full training set) ===  FPGrowth found 17 rules (displaying top 10)  1. [answering=1]: 64 ==> [query=1]: 62 <conf:(0.97)> lift:(5.57) lev:(0.01) conv:(17.62)  2. [database=1, oriented=1]: 96 ==> [object=1]: 93 <conf:(0.97)> lift:(18.06) lev:(0.01) conv:(22.71)  3. [materialized=1]: 56 ==> [view=1]: 51 <conf:(0.91)> lift:(33.21) lev:(0.01) conv:(9.08)  4. [concurrency=1]: 133 ==> [control=1]: 107 <conf:(0.8)> lift:(20.73) lev:(0.01) conv:(4.73)  5. [oriented=1]: 183 ==> [object=1]: 141 <conf:(0.77)> lift:(14.37) lev:(0.01) conv:(4.03)  6. [warehouse=1]: 73 ==> [data=1]: 56 <conf:(0.77)> lift:(7.26) lev:(0) conv:(3.63)  7. [expansion=1]: 84 ==> [query=1]: 60 <conf:(0.71)> lift:(4.11) lev:(0) conv:(2.78)  8. [processing=1, efficient=1]: 70 ==> [query=1]: 50 <conf:(0.71)> lift:(4.11) lev:(0) conv:(2.75)  9. [object=1, oriented=1]: 141 ==> [database=1]: 93 <conf:(0.66)> lift:(5.58) lev:(0.01) conv:(2.54)  10. [database=1, object=1]: 154 ==> [oriented=1]: 93 <conf:(0.6)> lift:(32.49) lev:(0.01) conv:(2.44) |

Step 4: (See code and txt files)

Note: I sort the result based on the purity of topic.

Step 5: (Nothing)